Why all this sudden attention on the Linux scheduler?

Amit Kucheria, PMWG Tech Lead
&
several people in the room

Linaro CONNECT

Code

```
(kernel/sched) $ wc -l core.c fair.c rt.c
 deadline.c idle_task.c stop_task.c
  8755 core.c
  6174 fair.c
  2094 rt.c
  1658 deadline.c
    98 idle_task.c
   128 stop_task.c
 18907 total
```

Which scheduler?

Completely Fair Scheduler (fair)
Realtime (rt)
Earliest deadline first (deadline)
IDLE (idle_task)
STOP (stop_task)

Problem

Throughput

Determinism

Throughput

Determinism

Solution

Throughput

Determinism
Power-efficiency

Determinism

Determinism: Problems

- Preemption: interrupts, locking
- Latency (interrupt -> processing, time between two consecutive runs of a task)
- Scheduling overhead

Determinism: Solutions

- Preemption: interrupts, locking
- Latency (interrupt -> processing, time between two consecutive runs of a task)
- Scheduling overhead

PREEMPT RT

ADAPTIVE NO_HZ

DEADLINE

Determinism: Features

Feature	**PREEMPT RT**	**ADAPTIVE NO_HZ**	**DEADLINE**
Physical process isolation*	No	No	No
Temporal Isolation	Yes#	Yes+	Yes
No scheduling overhead	No	Yes	No
Firm/Hard Real-time	Yes	No	No
Complexity	High	Low	Low

* Use cgroups + cpusets

With some limitations

+ Limitation of one task per core currently, else NO

Determinism

Requirements?

Power-efficiency

Power-efficiency: History

- ~~sched_mc~~
- big.LITTLE GTS patches (ARM)
- Packing Small Tasks (Linaro/ARM)
- Power aware scheduling (Intel)

Power-efficiency: History

- ~~sched_mc~~
- big.LITTLE GTS patches (ARM)
- Packing Small Tasks (Linaro/ARM)
- Power aware scheduling (Intel)

And then...

Ingo strikes

31st May 2013, Ingo Molnar on LKML:

"Today the power saving landscape is fragmented and sad: we just randomly interface scheduler task packing changes with some idle policy (and cpufreq policy), which might or might not combine correctly."

....

"_All_ policy, all metrics, all averaging should happen at the scheduler power saving level, in a single place, and then the scheduler should directly drive the new low level idle state driver mechanism."

...

"This is a "line in the sand", a 'must have' design property for any scheduler power saving patches to be acceptable - and I'm NAK-ing incomplete approaches that don't solve the root design cause of our power saving troubles..."

Power-efficiency: Proposal

Separate process and power scheduler (ARM)

Power-efficiency: Proposal

Separate process and power scheduler (ARM)

Topology
Idle + DVFS
Thermal

Acknowledgements

- **LKML**
- Vincent Guittot (Linaro/ST Micro)
- Morten Rasmussen (ARM)
- Catalin Marinas (ARM)
- James King (Linaro/Broadcom)
- Tuukka Tikkanen (Linaro/HiSilicon)
- Mike Holmes (Linaro/LSI)
- Charles Garcia-Tobin (ARM)
- Kevin Hilman (Linaro)
- Viresh Kumar (Linaro/ARM)
- Daniel Lezcano (Linaro)
- Others I've forgotten (apologies)

Linaro

CONNECT

Linaro

CONNECT